基于粗糙集聚类的高维离群点数据挖掘算法

文/陈　艳　朱建平

摘 要：本文回顾了离群点数据挖掘技术的研究成果，对高维离群点数据挖掘技术进行了简要的评述，并在此基础上提出了一种基于粗糙聚类的高维离群点数据挖掘的算法，采用粗糙集方法选择出最优子空间，直接对高维空间数据进行聚类，并对子空间离群点进行数据挖掘，取得了良好的效果。

关键词：离群点；数据挖掘；粗糙集聚类；高维空间

一、引言

离群点或异常点目前尚未有一个普遍认可的定义。Hawkins(1980)揭示了离群点的本质：‘离群点的表现与其他点如此不同，以至于让人怀疑它是由完全不同的机制产生的。”总之，离群点是指在大型数据集中，存在着不遵循数据模型的普遍行为的，与其余数据有很大的不一致的数据，其通常是由于测量误差，或是数据固有的可变性的结果。在现实社会中，离群点数据一方面伴有大量的噪声，另一方面又可能包括极有价值的信息，而大部分数据挖掘方法都将离群点数据视为噪声或异常情况而丢弃。

在实践中，离群事件往往比正常事件更令人感兴趣。在一些应用中，如金融保险领域的保险欺诈、信用卡欺诈、异常交易行为的分析；通信领域的手机盗打、网络入侵的检测；医学领域的病症诊断和药物分析等，离群数据挖掘技术都起着非常重要的作用。随着社会的不断发展，经济社会的各项活动越来越复杂，其中往往夹杂着大量的高维离群数据，这些离群点在多个方面、多个指标出现了异常，产生的行为可能带来更为严重的后果，因此，高维空间下的离群点挖掘更具有实用性。

二、粗糙集理论概述

粗糙集理论是由 Pawlak(1982)提出的一种用于处理不确定性和不精确性知识的数学工具。其基本思想是在保持分类能力不变的前提下，通过知识约简，提取分类或决策规则。它反映了人们处理不分明问题的常规性，即以不完全信息或知识去处理一些不分明现象的能力，或依据观察、度量到的某些不精确的结果而进行分类数据的能力。

粗糙集理论是基于信息系统所进行的研究，信息系统是研究对象的集合，而这些对象是通过指定属性和它们的属性值来描述的。信息系统表示为 $S=(U,Q,V,f)$，其中 U 是对象的非空有限集合，称为论域；$Q=C\cup D$ 是属性集合，子集 C 和 D 分别称为条件属性集和决策属性集，$V=\cup_{q\in Q}V_q$ 是属性值的集合，V_q 表示属性 $q\in Q$ 的值域，$f:U\times Q\to V$ 是 $U\times Q$ 到 V 的一个映射。

在粗糙集中，论域的划分是依据等价关系进行的。在信息系统 $S=(U,Q,V,f)$ 的任一子集 $A\subseteq Q$ 上，论域 U 上的一个等价关系为：

$$IND(A)=\{(x_i,x_j)\in U\times U:f(x_i,q)=f(x_j,q),q\in A \quad (1)$$

也记为 R_A，R_A 满足自反性、对称性和传递性，它是论域划分的一个准则。在信息系统中，所有等价关系的集合称为等价关系族。

约简是粗糙集进行数据挖掘的重要方法。它是在保持数据集原有的依赖关系及分类能力不变的前提下，利用数据集的等价关系删除冗余属性，对数据集进行简化，以获取知识。属性约简的目标就是要从条件属性集合中发现部分必要的条件属性，使得根据这部分条件属性和所有条件属性相对于决策属性有相同的分类能力，即约简前后所形成的相对于决策属性的分类一致。

在信息系统 S 中，若对于 $A\subseteq Q$ 满足 $R_A=R_C$，而且对于 $\forall q\in A$，$R_{(A-\{q\})}\neq R_A$，则称 A 为信息系统 S 的一个约简。所有约简的交集称为核，记为 Core(C)。属性的约简并非唯一的，通常一个信息系统中可能包括许多约简，但信息系统中核是唯一的且为任何约简的子集。

粗糙集理论能够在保留关键信息的前提下，完全利用数据本身提供的信息，无须任何先验知识；表达不确定或不精确的知识、分析不一致信息，对不确定、不完整信息进行推理，来获取知识。特别是，在属性约简中，利用等价关系对论域的划分，本质上是将高维空间中的数据点集投影到低维子空间中，并且可以利用约简后的属性集直接对高维空间数据集进行聚类。这与我们求解高维空间离群点的思路是基本一致的，因此，我们选择粗糙集技术，借助一种基于信息量的属性约简算法，将高维空间数据投影到低维空间，进行粗糙集聚类，并在此基础上寻找离群点。

三、离群点数据挖掘的新方法

(一)粗糙集聚类

在粗糙集中，信息系统的核是唯一的，而且它是任意约简的子集。因此，我们可以将核作为求最小约简的起点，向属性子集中不断加入最重要的属性，直到终止条件满足。下文我们将给出信息熵和属性重要性的概念，并以此作为属性约简的终止条件。

信息熵通过事件发生的概率来表达知识所含的信息量，它度量了信源提供的平均信息量的大小，我们可以用其来表示信息系统中，等价关系所包含的信息量。在信息系统

基金项目：本文获国家教育部‘新世纪优秀人才支持计划”(NCET-04-0608)资助。

S=(U,Q,V,f) 中，已知等价类族 $U/R_A=\{X_1,X_2,\cdots,X_n\}$，$A\subseteq Q$，则 A 的信息熵的定义为：

$$H(A)=-\sum_{i=1}^{n}p_i\ln(p_i) \quad (2)$$

其中 $p_i=|X_i|/|U|$ 为等价类 X_i 的概率，$|X_i|$ 为 X_i 的基数，$|U|$ 为 U 的基数，$|U|=\sum_{i=1}^{n}|X_i|$。

属性的重要性是用来表征某一属性的重要程度，它用添加或删除某个属性所引起的条件信息量的变化大小来度量。在信息系统 S 中，$A\subset C$，则对于 $\forall a\in(C-A)$ 在 A 中的重要性定义为：

$$SGF(a,A,D)=H(D|A)-H(D|A\cup\{a\}) \quad (3)$$

由定义可知，当 $A=\emptyset$ 时，$SGF(a,A,D)=H(D)-H(D|a)$，$SGF(a,A,D)\geq 0$，且 $SGF(a,A,D)$ 的值越大，在已知等价关系 R_A 的条件下，属性 a 对于决策属性 D 越重要，必须将该属性加入到约简属性子集中。当 $SGF(a,A,D)=0$ 时，属性约简终止。

下面给出了基于信息量的属性约简算法，来求解最优子空间。

(1) 确定信息系统的核。首先设属性 A 集为空集，即令 $A=\emptyset$，计算信息系统 S 中的条件信息熵 $H(D|C)$、各属性条件信息熵 $H(D|a_i)$ 及各属性的重要性 $SGF(a_i,A,D)$，其中 $a_i\in C$。若 $SGF(a_i,A,D)>0$，则 a_i 为核属性，加入核 Core(A) 中，即：

$Core(A)=\{a\in A|SGF(a,A,D)>0\}$。

(2) 终止条件的初始判断。一般情况下，核只是最小约简的一个子集，其信息量小于条件属性集的信息量。令 $B=Core(A)$，如果 $H(D|B)=H(D|C)$，则核为信息系统的最小约简；否则，以核 Core(A) 为起点，加入新的非核条件属性，直到满足终止条件为止。

(3) 寻找最小约简。求各条件属性信息熵 $H(D|B)\cup\{a_i\}$ 和各属性的重要性 $SGF(a_i,B,D)$，其中 $a_i\in(C-B)$。

选出重要性 $SGF(a_i,B,D)$ 最大的属性，将其加入到属性集中，即 $B=B\cup\{a_i\in(C-B)|\max SGF(a_i,B,D)\}$。如果有几个属性 $a_i\in(C-B)$ 具有相同的最大重要性时，则选择与 B 的复合属性值最少的属性加入属性集 B 中。

计算新的 $H(D|B)$，若 $H(D|B)=H(D|C)$，则属性约简终止，得到最小约简 B；否则，再加入新的属性，重复此步计算。

最小约简 B 就是保持原始数据集分类信息的最低维度空间的属性集，其所对应的等价关系 R_A 就是最优等价关系。因此，最小约简 B 的生成过程，就是高维数据子空间的生成过程，也是最优等价关系的生成过程。基于此最优等价关系及其相应属性集的所形成分类是最优且有效的分类。在粗糙集属性约简过程中，直接进行了聚类分析，即粗糙集聚类，得到最小属性约简和高维数据的子空间最优分类。

(二) 基于聚类的离群点数据挖掘

通过粗糙集聚类，我们可以将高维空间数据投影到低维空间，求得数据集的最优子空间及有效分类。下面我们将基于这一最优子空间和有效分类进行离群点数据挖掘。

Ramaswamy 等 (2000) 在基于距离的离群点定义的基础上，提出了一种基于距离的 k- 最近邻 (k-NN) 离群检测算法，根据对象的 k- 最近邻距离赋予每个对象一个离群点得分，从而来找出离群点。其基本思想是：如果一个数据总是远离大部分的数据点，则它就是离群点。按照这一思想，我们提出了一种基于聚类的离群点检测算法，把数据集中的每个记录看作是空间上的一个点，计算每两点之间的距离，并依据粗糙集聚类的结果，给出各分类的离群得分，找出与其它点相距距离最大的点或类，该点或类即为离群数据。

首先，给定三个参数，最小类的大小 (类中点的个数) n、近邻点的个数 k 和离群距离阀值 θ。

其次，对可能的离群点或离群类进行初步判断。根据粗糙集聚类的结果，比较各类的大小，如果某一类中点的个数小于最小类，即 $n_i<n$，则将此类初步判定为可能离群的类，等待进行进一步的分析。

再次，计算出各可能离群的类中所有点的 k 个最近邻距离。$D_{ik}(p)$ 表示第 i 个可能离群的类的点 p 和它的第 k 个最近邻的距离。本文以最常用的欧氏距离来表示空间中两数据点之间的距离，则任意两数据点间的距离为：

$$d(X,Y)=\sqrt{\sum_{i=1}^{r}(x_i-y_i)^2} \quad (4)$$

然后，计算各可能离群的类的离群得分。根据每个点的个 k 最近邻距离，计算点离群得分，定义为 k 个最近邻距离的平均数，记为：

$$score_i(p)=\frac{1}{k}\sum_{j=1}^{k}D_{ij}(p) \quad (5)$$

同时，计算出各可能离群类的类离群得分，表示为：

$$score_i=\frac{1}{n_i}\sum_{p=1}^{n_i}score_i(p) \quad (6)$$

最后，确定离群类及离群点。比较可能离群的类的离群得分 $score_i$ 与离群距离阀值 θ 的大小，将离群得分 $score_i$ 大于阀值 θ 的类及点作为高维空间数据集的最终离群点。

四、结束语

在高维空间离群数据挖掘方面，国内外已有不少的研究成果，主要方法都是基于把高维空间数据投影到低维子空间，在最优子空间上进行挖掘这一思想，本文沿袭了这一思路。但与以往的方法不同之处在于，本文采用粗糙集中属性约简的方法，进行粗糙集聚类，求解出最优子空间，并采用基于聚类的离群点数据挖掘方法，通过离群得分来发现高维空间数据集中的离群点。

本文所提出的算法能够较好进行高维离群点检测，但在参数值 k、n 和 θ 的选取，以及算法性能的提高等方面还有待改进。同时，本文由于数据资料的局限，仅对高维空间离群数据挖掘算法进行了定性分析，在定量方面还需进一步的研究。

参考文献：

[1] 黄洪宇，林甲祥，陈崇成等. 离群数据挖掘综述 [J]. 计算机应用研究. 2006(8): 8- 13.

[2] 魏藜，宫学庆，钱卫宁等. 高维空间中的离群点发现 [J]. 软件学报. 2002(2): 280- 290.

[3] Z Pawlak. Rough sets [J]. International Journal of Information and Computing Science, 1982(5): 341—356.

作者单位：厦门大学经济学院

(责任编辑：曾 鸿)